Data Management and AI in E-government*

Tibor Vámos and István Soós

Computer and Automation Research Institute, Hungarian Academy of Sciences
H-1111 Budapest, Lágymányosi u. 11., Hungary
vamos@sztaki.hu

Abstract. Similarities and differences between large companies and e-government data management are discussed. The major issue of e-government systems is the human face, i.e. the citizen–administration dialog. Openness vs. individual and administration procedure security are delicate legal and technological problems. A regional experiment is reviewed that applies three languages: natural language of the citizen, formal one of the administration and quasi-natural one of the dialog. Natural language understanding is a special feature using subject-oriented closed-world vocabularies and scenarios. The combination provides a decision support especially dedicated to the citizen.

1 Introduction and Motivation

Data management problems faced by e-government generally do not differ too much from similar issues in large companies. This statement is not without controversy; however it reflects the authors' views. A modern government should be run, in many ways, like a large, well-managed company. This implies that the basic system tools in use should be similar to well-developed commercial data management systems that are already available, and that it would be a waste of money to develop any special system that has not stood the test of time and does not have a well-developed support system.

Even so, there are some differences that need special attention. These are issues related to privacy and user interface. A democratic government is basically an open one: the Latin expression 'res publica'—originally meant 'public affair' **—means anything owned by the people referring in Roman Law to all public ownership and best expressed by the concept of SPQR – Senatus Populusque Romanus; and, contrary to the underlying needs in the management of a private business, secrecy should only be required under very exceptional circumstances. On the other hand, due to the fact

* The project was supported by NRDP Grant No. OM-00510/2001

** originally res publica (see Cicero: De re publica libri VI, I. (39): "Est igitur, inquit Africanus, res publica res populi, populus autem non omnis hominum coetus quoquo modo congregatus, sed coetus multitudinis iuris consensu et utilitatis communione sociatus. eius autem prima cause coeundi est non tam inbecillitas quam naturalis quaedam hominum quasi congregatio; non est enim singulare nec solivagum genus hoc, sed ita generatum ut ne in omnium quidem rerum affluen<tia>")

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-3-540-44836-5_33

M.A. Wimmer (Ed.): KMGov 2003, LNAI 2645, pp. 230–238, 2003.

that the governments, especially local authorities, deal with their citizens' individual problems, too, it is imperative that they abide by strict privacy principles. Privacy is not the same as security in commercial systems; it is a delicate web of interaction and decision-making communications rather than a simple system closure task as required in company management or the management of government organizations dealing with national security.

In the aftermath of September 11 we are seeing just how delicate priva-cy really is. It essentially comes down to a conflict among different valid interests. The other side of the problem is the social support agenda. Why a certain people should be the recipient of money collected as tax is an open question for a community; for the beneficiaries however it is their own private problem.

The other difference is the nature of the user interface. Commercial systems, such as banks have well-developed user communications; how-ever these are based on rather simple rule sequences leading to narrowly defined results. E-government, especially at the local level, deals with a wide variety of its citizens' individual problems. These citizens are, for the most part, unable to formalize their queries into a machine understandable, unambiguous text; and oftentimes they are unaccustomed to the use of regular questionnaires. However, the introduction of e-government should not increase the alienation of the citizens from public administration.

Our project is a local government experiment in a region with about 120.000 inhabitants; at its center there is a university city; there are also several rural settlements in the region at varying levels of economic and cultural development. 70 localities (city and villages) are included. The project has two special emphases: it addresses the e-government issue from a certain sociological perspective and it experiments with advanced artificial intelligence tools. The two are rather closely related: the first is built on human relations and the second is the creation of a human face through artificial intelligence, the flexibility of which allows for the reproduction of more sophisticated human attitudes by machines and the creation of a machine support for human activity. This last aspect is of particular significance: By the automation of routine services more time and attention can be devoted to cases that require human attention. The possibility of personal consultation with a sophisticated system prepares the citizen for better understanding of his/her chances, as well as for the presentation of his/her case to the administration.

The major contribution of AI to the human face of machine-supported administration is the human-oriented dialog, i.e. the understanding of natu-ral language. It can serve as a tool for real dialog that not only checks whether both partners, i.e. the administration and the citizen, understand each other, but also helps them in the process. The citizen not only receives the resolution of his/her case, but can follow each step of the procedure, get all related regulations from the original legal text, as well as, comments and interpretations, all as a part of the dialog and in a language closer to his/her linguistic usage.

How the communication takes place is another human issue. Currently the citizen has to go to the administration office, in most cases several times, wait in humiliating conditions and spend his/her own precious time. The process itself can take an unreasonable amount of time, and be both an irritating and costly experience (in terms of economic loss) for the citizen. With intelligent machine-supported administration this all can be done from home, at any time of day and, possibly, in one short process.

2 Interfaces

The interface problem here is very different from the problems company wide systems face. A well-designed company system has uniform inter-faces and mostly standardized internal representations. For practical reasons this cannot be achieved in public administration. Different branches of government and of local administrative boards have different standards; they have some autonomy in choosing the software they use. All have their specific history, and receive their financial backing at different times. The platforms, operating systems, basic software products in use and procedural customs are all somewhat different; the unification would be a very costly investment, unrealistic in most cases; and it would be met with much resistance. This entails the need to develop practical standardized interfaces and gateways, and to investigate the possibility of legislating the use of the interface system in administrative practice.

The other interface problem is privacy and data protection. Existing laws forbid all forms of data unification, even within a same administrative board. Data unification means the collection of different personal data in the same file, e.g. taxation data with health records. This is regulated to prevent the different bodies of administration from getting an overall picture of the citizen that could then be used to intrude in his/her private life. The unification can only take place with the informed consent of the data owner, for whom even the context of the retrieval and transfer of the data should be accessible. Another directive requires either the deletion of the transferred data once the case has been settled, or the preservation of those in line with specific laws. Statistical surveys require special attention notably with regard to data purification of different kinds of statistics. The gateway service would need to have an automated system for the registration of all transactions to ensure data is not misused. This is a far more complex task than the sensitive controls of banking systems.

3 Languages

Our system uses three different versions (vernaculars) of the same language. The first is the language of the citizen, mostly poorly educated, unable to present his/her case in a concise, unambiguous and non-redundant form. This unfortunate fact is supported by the recent European review of literacy-related understanding and a similar study, a few years earlier, in the US. On the other hand, the language used in legal texts is difficult for any-one who does not have a background in law to understand, due to the strong requirements for a professional logical definition power. This is the second language, the formal language of the administration. Although this is a rather well formulated procedural communication tool based on legal practice, further refinement is required to make it into a strictly logical description of the case, avoiding all ambiguity; or, when ambiguity is inherently present in the case, to incorporate probabilistic logic into the description. This language is a schematic description of the procedural scenarios.

The third language is for communication with the citizen. Basically it is a popularization of the formal language using everyday expressions for a more user-friendly communication. Its first role is a retranslation of the citizen's claim, to check

if the system understood it well and the applicant agrees with the interpretation. It is a more advanced form of Feigenbaum's Eliza dialogue. Its second role is to query for missing information. The third role is the communication of the result of the administrative procedure. This last role, when needed, must be both empathetic and convincing. Remember what the humorist George Mikes wrote about a letter from a British civil servant in his How to be an Alien!

4 Special Databases

The system has two kinds of special databases. The first contains specific vocabularies. These are semi-automatically selected from claim documents that the administration has handled in the past. The words are selected ac-cording to their relevance to the subject, stemmed and grammatically analyzed by the HUMORESK software of MORPHOLOGIC. A large Hungarian Language Corpus is used to further analyze the words. This Corpus was developed by the Szeged University in cooperation with MORPHOLOGIC. The Corpus also helps in the selection of non-regular words, especially names of persons and companies that are important for the understanding of the claim.

The vocabular2ies are broken down into different administrative subjects; this is the basis for direction and refinement in the interpretation. The text to be interpreted is analyzed by the weight of relevant, preselected words as stored in the subject-oriented vocabularies. The recognition of simple grammatical relations, such as the who's what genitive noun relation and the who what accusative verbal phrase, further helps in the understanding of the context. The vocabularies contain a thesaurus of synonyms, based on the earlier claim documents. This is of particular interest for a further socio-linguistic analysis.

5 Understanding, Case-Based Reasoning

We stop here for a moment. One of the main philosophical issues in discussions about AI is the concept of understanding. For us understanding means the perception or identification of a piece of information as it relates to con-sequences. In most cases, and the one we are dealing with here is similar, the understanding of a given piece of information entails the triggering of a specific response scenario. This is the role of our second category of data-bases.

The second category of databases contains the possible scenarios for regular claims and the related administrative actions. These scenarios are rather general, the head of the scenario graph is always the naming of the subject (social relief, complaint against somebody or some organization, construction permission, etc.) and branches lower down, as well as the nodes are also all well-definable, and follow the logical steps used at the given administrative department. The terminal expressions of the graph will often be vague, here the thesaurus and the question-answer dialogue are helpful.

The procedure is a typical case-based reasoning, combined, if possible with rule-based processes. The project, in putting its emphasis on better meeting human needs, puts its emphasis on the case-based problems; these are also more attractive from the

point of view researchers in artificial intelligence. We have already experimented with two different subjects: children custody decisions after a divorce, and case-study support in US bank-ruptcy law.

Technically the task is the matching of the two databases. Given how the administrative subjects are well defined and have limited vocabularies, this has so far proved to be less difficult than the understanding of discretional texts. Most goal-oriented human communication is subject-limited and the professional procedures create closed linguistic entities. The real problem is the translation of the natural language of communication used by the unprepared citizen.

The basic trick for a workable system is the limitation of linguistic components. The analysis of hundreds of texts provided by the local authority showed this to be feasible. No specific subject used more than 2000 relevant words. By relevance we mean relevant from the point of view of under-standing. Filtering for relevance was done using word statistics and manual extraction. The vocabularies contain several vernacular-specific synonyms but the number of these is not too significant either. Although they change in time with the changes in linguistic fashion, their usage is more uniform and poorer than any sophisticated discourse. To account for this phenomenon, a learning routine is added. With official legal texts allowing no room for ambiguity, their vocabulary is also limited; it *must* reflect a closed conceptional world.

However, in the case of non-understanding by the system, the human dialog is helpful and this is the point where the machine process should stop and switch over to human communication. The machine filters all seemingly regular cases and helps the human expert concentrate on the truly problematic cases. This is the other contribution to the more human face of the machine-supported governance – as was earlier emphasized.

As was mentioned above, the nature of natural language allows for an interesting experiment. The language used by the claimant is full of information about the true status or nature of that claimant. The subjective psycho-logical impression that a skilled administrator will have of a given claimant will often be to the claimant's advantage; of course, this psychological impression can just as easily be the basis for unwanted, incorrect prejudices. A socio-linguistic analysis here is an attempt to help understand the emphasis needed for a given case all the while avoiding preconceptions.

Two simplified examples show how the much more complex systems work (see Annex).

The experiment is ongoing and is progressing well. A thorough information flow analysis of the whole administration, including all the departments of the self-government has been completed. Figures 1 and 2 show examples of information flow diagrams. Each procedure was analyzed in terms of discrete actions; on average there were 30 to 50, but some procedures required 100 discrete actions. The analysis included the active time of process handling, and the waiting times for delivery within the administration, request for data and other idle periods. The result was a factor of more than 100; the active process time of manual handling was a few hours, while the whole process lasted several weeks. This means that a complete electronic dialog and data acquisition process among the different administrative units can reduce the process time to an hour or less instead of long weeks.

The system would naturally be expected to handle voting and opinion polls, as well as inquiries related to the operation of the governance and chat groups on these subjects. The main software frame is an interactive portal including all usual services.

The work is based on written texts; technology for vocal understanding, as would be needed to understand the wide variety of utterances and texts, has yet to be developed.

The system is partly under installation. There are several vandal-proof terminals in units similar to telephone booths providing open access for less sophisticated procedures, while other terminals are located in telehouses, schools and libraries and come with supporting personnel. Participants in this experiment working with the local administration all have a terminal at home.

The first experiments, after successful laboratory tests, will be orga-nized between the system's designers and a group of local administration personnel. The laboratory tests are ongoing, the project as an experiment should be completed by the end of the first quarter of 2004.

The evaluation of the project has two levels. On the one level the socio-logical group of the Budapest University of Economy and State Administration is working on the second survey and they will close the project with a third one. On the next level the founding National Research Development Project regularly supervises the progress.

In its first phase the experiment is to provide support to the administration and get feedback from the citizens.

6 Conclusion

The feasibility of an active data management system for e-government is demonstrated using advanced artificial intelligence. The experiment is carried out in a local government setting; i.e. an environment in which there is a significant amount of human contact involving direct, natural language communication with individual citizens. Certain legal and technical problems also arise; but they can be resolved.

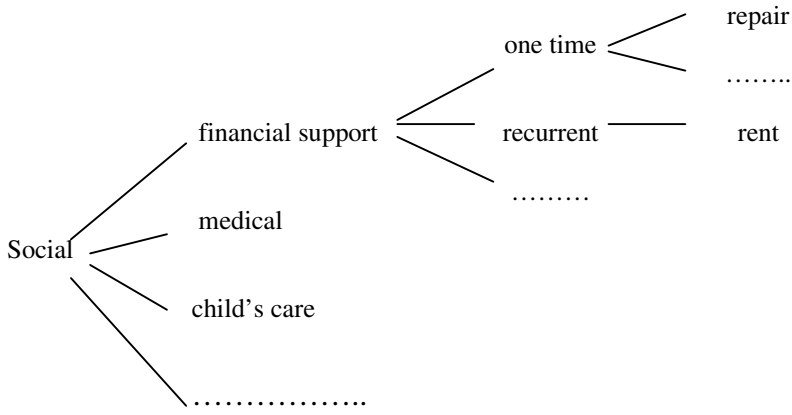
Annex

Example 1

"Dear Mr. Major,
Forgive me for disturbing you and for coming to you with my personal problem but I am unable to help myself. We have no cash; we are disabled *pensioners*. Our pension is 15000 + 5000 Ft and we cannot pay the *rent* that the owner of our very humble home wants. We need some help…"

This is an extract from a longer letter to Mr. Major. The words in italics clearly show that this is a social problem; the claimants would like to get some regular money (rent) for their home or some other solution for their dwelling.

The system identifies the nature of the problem and tries to match it with a scenario in the case base:



The structure is naturally more complex; the scenario-base contains all typical social problems.

The grammatical analysis identifies the relations: How many persons are involved, what is the relation (mostly family) of the persons, who owns what, who is in debt to whom, etc.?

The redundancy of the text and the use of some typical words, e.g. cash for money are characteristic for the claimant. The style suggested by the words and grammar used by the claimant also indicates his/her age.

If the scenario is clear, the translation is given with a related text, in this case:

'We understand that you cannot pay the rent for your home. Do you need any regular support above and beyond your disability pension or would you like to move into a social home?'

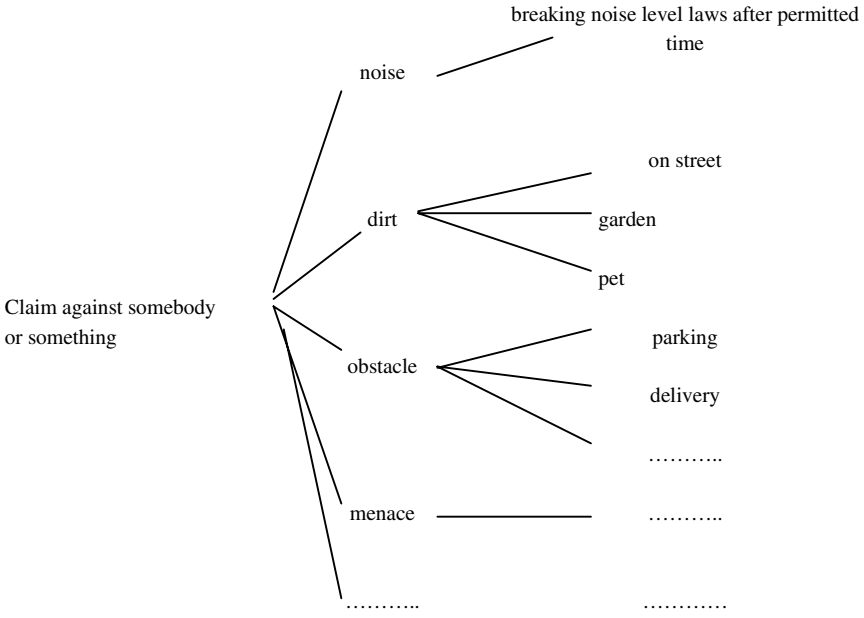
After clarification the final answer is also attached to the scenario:

'Unfortunately we are unable to raise your retirement pension due to both the current pension regulations and the budget of the community. On the other hand, you have now been added to a waiting list for social accommodation that is within your financial means. The expected waiting time is about two years. To meet your immediate needs you will receive a one-time extraordinary support of 2000 Ft.'

Example 2

"We *cannot sleep* because of the *horrible* people making *noise* under our window. It is being done by the *Yellow Submarine Disco* and it goes on *long in the night* and they leave all their *trash out* in the morning…"

The scenario is simple:



'Do you mean that the Yellow Submarine Disco breaks the noise level laws and that they leave their garbage on the street?'

Answer of the system:

'We will send the local police after 10 pm to measure the noise levels, and after 11 pm to check the levels again. We will ask the city's garbage collection company for a report. Following their report we will take measures as prescribed by our regulations and inform you about the result.'

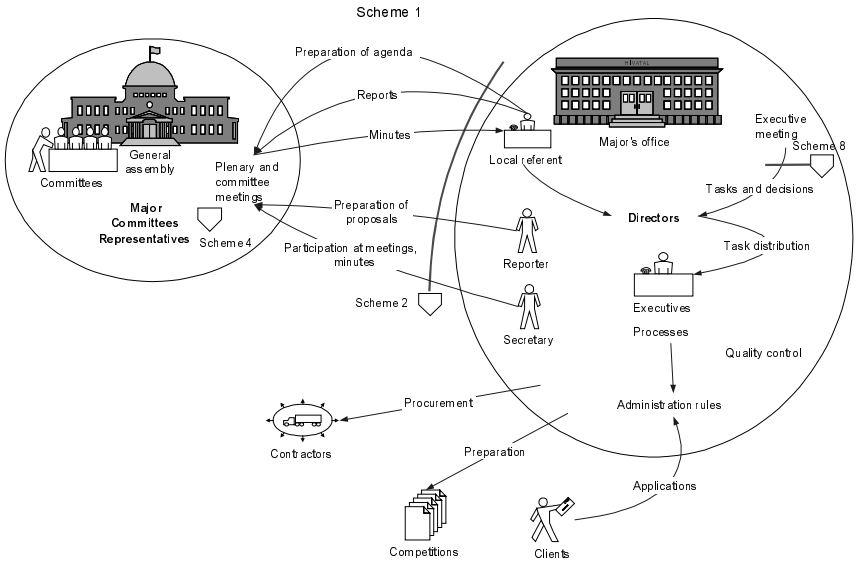


Fig. 1. General model of the self-government activity

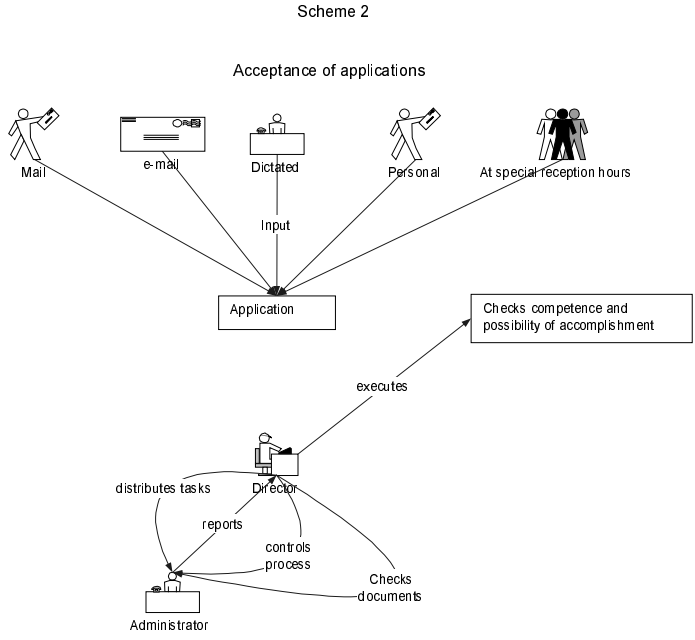


Fig. 2. General rules